Journal of Commerce & Accounting Research
9 (4) 2020, 73-77
http://publishingindia.com/jcar/

Submitted: 10 August, 2020
Accepted: 28 October, 2020

DRIVERS OF HOUSING PRICES IN THE GROWING, SUBURBAN SOUTH OF THE US: EVIDENCE FROM CUMMING, GEORGIA

Mitra L. Devkota*, Eric B. Howington**

Abstract *This study examines the determinants of housing prices for single family homes in Cumming, Georgia, USA. Cumming is located with Forsyth County, one of the fastest-growing suburban areas in the United States. Results of the regression analysis indicate that the numbers of bathrooms, square footage, and school zoning are all statistically significant predictors of the price of a single-family home.*

Keywords: *Regression, Zillow.com, Housing Prices, Multicollinearity, Residuals, Georgia*

INTRODUCTION

"2020 Best Counties for Families" (2020) rates Forsyth County, Georgia, a rapidly growing county located just north of Atlanta, as the second best county for families in America. Forsyth County is representative of many southern suburban counties experiencing a significant growth in recent years. According to the U.S. Census Bureau, Forsyth County ranked ninth in the United States in percentage growth over the years 2010-2018 with a percent growth of 34.8% ("New Census Bureau Estimates," 2019). Cumming, Georgia is the county seat of Forsyth County. At the 2010 Census, Cumming had a population of 5,430, but there are approximately 100,000 Cumming mailing addresses when the surrounding unincorporated areas are considered. Naturally, rapid growth generates real estate activity, an increased number of property transactions, and, often, increased sales prices. What home characteristics are the most significant drivers of price for single-family homes in this rapidly growing southern town?

This paper analyzes determinants of real estate prices in Cumming, Georgia. We analyze the relationships between the home characteristics and prices using cross-sectional data for single-family houses in the year 2020. It is expected that the findings of this study will help home-buyers, sellers, real estate agents, and researchers better understand the relationship between these drivers and housing prices of this city and similar growing, southern, suburban areas. The remainder of the paper is organized as follows. A review of previous empirical studies is carried out in section 2. A detailed description of the data, variables, and methodology used in the study are presented in section 3. The results and findings are presented in section 4. Section 5 concludes the paper and discusses the results.

LITERATURE REVIEW

There are papers in the literature analyzing housing prices in other countries; however, we are not aware of any published literature that has analyzed the drivers of housing prices for rapidly growing southern US towns like Cumming, GA. Kain and Quigley (1970) estimated the market value of specific aspects of the bundles of residential services consumed by urban households. Regressing the market price of owner and renter-occupied housing units on several quantitative and categorical predictors, the authors found that the housing prices are significantly affected by the number of bedrooms, number of bathrooms, and size of the house. As expected, the study also confirms the influence of the neighborhood schools on housing prices. Similarly, Cebula (2009) found that housing prices in Savannah, Georgia, were influenced by the number of bedrooms, the number of bathrooms, and the presence of swimming pools. There are also a number of published studies from other countries. A study by Selim (2009) found that housing prices in Turkey were significantly influenced by the number of bedrooms, number of bathrooms, the size of the house, the location, and the presence of swimming pools. Randeniya, Ranasinghe and Amarawickrama (2017) studied the house price of data for Sri Lanka using 50 single-house transactions in Maharagama urban neighborhood area to illustrate the applicability of the hedonic pricing model. The authors used correlation analysis to study the degree of relationship between the variables. The results suggest that design type of the house, distance to

[*] Mike Cottrell College of Business, University of North Georgia, Dahlonega, GA 30597, USA. Email: mldevkota@ung.edu
[**] Langdale College of Business, Valdosta State University, Valdosta, GA 31698, USA. Email: ebhowington@valdosta.edu

the local road, quality of infrastructure, garden size, number of the bedrooms, and age of the house were the determining factors to estimate the price of the house. Bui (2020) investigated the determinants of apartment prices in Ho Chi Minh City of Vietnam. Using multiple regression analysis on survey data of 124 apartments traded during the first six months of 2019, the author found that the apartment price was positively related to size of the apartment, presence of balcony, presence of swimming pool, presence of shopping malls, and periodic rental income, and negatively related to the distance from the city center. Calmasur (2016) used the hedonic pricing model to determine the factors that influence housing prices in Turkey. They used 1200 data points in their study, and found that the number of bedrooms, presence of shopping malls, and other included facilities had impact on housing prices.

Recently, real estate pricing methods have come under scrutiny (Belke & Keil, 2017). The multiple regression analysis is a popular tool for appraising single-family houses. Dublin (1998) showed how the correlations between the prices of the neighboring houses could be incorporated in estimating regression coefficients and in predicting housing prices. Using multiple listings data from Baltimore (USA), the author also discussed the practical difficulties inherent in using kriging to predict the housing prices. Newsome and Zeitz (1992) demonstrate how the problem of heteroscedasticity can be minimized by developing separate multiple regression models for houses in different prices. They further mention that their technique could be applicable in the valuation of commercial properties as well. Similarly, Yusof and Ismail (2012) used the multiple regression analysis to explain price variation for houses in Malaysia. The study found that locality and the area of the house were the most influential factors in determining the price of houses. When using multiple linear regression, housing price appraisers should be aware of some sources of statistical distortion such as nonlinearity, multicollinearity, and heteroscedasticity.

DATA AND METHODOLOGY

The data set used in this study consists of a sample of 92 single-family homes in Cumming, Georgia, in the year 2020. This data set was collected from real estate research site www.zillow.com[1]. Our interest is to estimate the listing price of the house using the predictors number of bedrooms, number of bathrooms, square footage, lot size, age, number of parking spaces, whether the house is in active listing or not, and the high school for which the house is zoned. These variables are described in Table 1 below:

Table 1: Variables used in the Study

Variable	Description	Details
Price	Listing price of the houses	Ranges from $189,000 to $689,000
Bedroom	Number of bedrooms	Ranges from 2 to 8
Bathroom	Number of bathrooms	Ranges from 2 to 6
Sqft	Floor size in square feet	Ranges from 1322 sq. ft. to 6165 sq. ft.
Lot size	Lot size in square feet	Ranges from 0.01 acres to 2.69 acres
Age	Age in years	Ranges from 0 to 52
Parking	Number of parking spaces	Ranges from 1 to 6
Active	Whether the house is in active listing	Yes/No (Two categories)
High School	The name of the high school assigned	Name of High Schools (Six categories)

Here, "Active" is a categorical variable with two categories—"Yes" for active on the market for sale and "No" for pending or already sold. Similarly, "High school" is a categorical predictor with six categories: Forsyth Central High School (Central), Denmark High School (Denmark), Lambert High School (Lambert), North Forsyth High School (North), South Forsyth High School (South), and West Forsyth High School (West). These categorical variables have been re-coded as 0-1 dummy variables. Active [Yes] (meaning that the house is in active listing in the market) is used as the reference category for Active. Thus, the dummy variable "Active" represents the difference between the price of the homes that are in active listing and otherwise. Similarly, West Forsyth High School is used as the reference category for High School. Thus, the dummy variable "Central" represents the difference in price between a home in Forsyth Central High School and West Forsyth High School, the dummy variable "Denmark" represents the difference in price between a home in Denmark High School and West Forsyth High School, and so on. Following Pardoe (2008), half bathrooms are valued 0.1 (meaning that two full bathrooms and one half bath is considered 2.1 bathrooms) to reflect the belief that half bathrooms (one with toilet and sink only) are not valued by home-buyers nearly as much as full bathrooms.

The number of bedrooms is a quantitative variable. However, due to the small number of unique values for number of bedrooms, a scatterplot may not be the best way to examine the relationship between the prices of the homes and the number of bedrooms. The same logic applies for the number of bathrooms and the number of parking spaces, so we used the side-by-side boxplots of price against these three variables to examine the relationship between them. The plots (not shown, but available upon request to the authors) show an approximate general tendency of increase

[1] Zillow.com, founded in 2005 by Rich Barton and Lloyd Frink, is one of the most-visited U.S. real estate sites on the Web (DeVeaux & Velleman, 2012).

in price with an increase in the number of bedrooms, the number of bathrooms, and the number of parking spaces, with the exception of four parking spaces. Additionally, the scatterplots of price against age, square footage, and lot size show approximately linear relationships between the pairs of the variables. Thus, it is reasonable to fit a multiple linear regression model to estimate the price of a house (response) using the number of bedrooms, the number of bathrooms, square footage, lot size, the number of parking spaces, status of the house (active or not) and the High School as the predictors. The desired multiple linear regression model is

$$price = \beta_0 + \beta_1\, Bedroom + \beta_2\, Bathroom + \beta_3\, sqft + \beta_4\, Lotsize + \beta_5\, Age + \beta_6\, Parking + \beta_7\, Active + \beta_8\, Central + \beta_9\, Denmark + \beta_{10}\, Lambert + \beta_{11}\, North + \beta_{12}\, South + \in \qquad (1)$$

where β0, β1,....,β12 are the regression parameters to be estimated, and e is the random error term.

According to Wooldridge (2011), multicollinearity is possible if the t-statistics corresponding to the parameter estimates of independent variables in a linear regression model are not statistically significant, whereas the overall F-statistic is statistically significant. They further add that multicollinearity is a serious problem if the variance inflation factor (VIF) is greater than 10. In this regard, a linear regression model is fitted using price of houses as a response variable and the number of bedrooms, number of bathrooms, square footage, lot size, and number of garages as predictors, and the results are reported in Table 3. The VIFs for all the variables are smaller than 10 indicating that multicollinearity is not an issue in the multiple linear regression of price on the number of bedrooms, number of bathrooms, square footage, lot size, number of garages, status of the house (active or not), and the high schools assigned to the houses.

The estimated multiple linear regression equation is

$$\widehat{price} = 135582.46 + 14640.57 Bedroom + 22233.07 Bathroom + 31.50 sqft + 36113.6 Lotsize - 1243.09 Age + 8102.45 Parking - 16164.17 Active - 15791.96 Central + 5544.18 Denmark + 78033.97 Lambert - 56409.39 North + 30122.89 South \qquad (2)$$

RESULTS AND FINDINGS

We have created a correlation matrix to examine the strength of linear relationships between the variables. The correlation matrix is displayed in Table 2. The figures indicate that there are positive linear associations between the price and bedrooms, price and bathrooms, price and square footage, price and lot size, price and number of garages, and a negative linear association between the price and the age of the house. We also observe that there is a very strong correlation between the bedrooms and the square footage, and between bathrooms and square footage. This suggests that multicollinearity might be an issue between these pairs of variables (Anderson, Shoesmith, Sweeney, Anderson & Williams 2014; Bowerman, O'Connell & Murphree 2014).

Table 2: Correlation Matrix

	Price	Bedroom	Bathroom	Sqft	LotSize	Age	Parking
Price	1.00	0.67	0.69	0.78	0.09	-0.17	0.34
Bedroom	0.67	1.00	0.70	0.78	0.08	-0.07	0.24
Bathroom	0.69	0.70	1.00	0.79	-0.01	-0.25	0.15
Sqft	0.78	0.78	0.79	1.00	0.13	-0.17	0.31
LotSize	0.09	0.08	-0.01	0.13	1.00	0.43	0.17
Age	-0.17	-0.07	-0.25	-0.17	0.43	1.00	0.08
Parking	0.34	0.24	0.15	0.31	0.17	0.08	1.00

The real estate data were sampled from a larger set of records for sales in the year 2020. Thus, we believe that the data are representative of a particular population – all the houses for sale in Cumming, Georgia. The error the model makes in predicting one price is not likely to be related to the error it makes in predicting another. In assessing the fit of the model, the residual plot shows no pattern, and that the errors are roughly normally distributed and have mean zero. Additionally, a plot of the residuals against the fitted values exhibits no serious violation of the constant variance assumption. Thus, the assumptions of the multiple linear regression model (1) appear to be satisfied. The regression model is statistically significant (F = 25.82, p < .0001) for any conventional level of significance. Additionally, this model is the one selected by best subsets regression criteria (as the model has the highest value of adjusted R square and the lowest value of AIC). Nevertheless, the predictors such as bedroom, garage, and the dummy variables associated with two of the high schools-Central and Denmark-are not significant. The coefficient of determination of our regression model is 0.7532. This means that 75.32% of the variation in the prices of the houses is accounted for by our regression model.

Table 3: Parameter Estimates Table for Multiple Regression Model

Parameter Estimates					
Term	**Estimate**	**SE**	**t-Ratio**	**Prob>|t|**	**VIF**
Intercept	135582.46	37933.40	3.57	0.00***	.
Bedroom	14640.57	9987.89	1.47	0.15	2.84
Bathroom	22233.07	11046.52	2.01	0.05**	2.99
Sqft	31.50	12.05	2.61	0.01***	4.63
LotSize	36113.60	19708.35	1.83	0.07*	1.76
Age	-1243.09	751.55	-1.65	0.10*	1.58
Parking	8102.45	8290.32	0.98	0.33	1.27
Active[No]	-16164.17	8641.30	-1.87	0.07*	1.10
High School[Central]	-15791.96	14982.34	-1.05	0.30	1.81
High School[Denmark]	5544.18	13781.73	0.40	0.69	1.67
High School[Lambert]	78033.97	17187.01	4.54	<.0001***	1.97
High School[North]	-56409.39	16724.21	-3.37	0.00***	2.32
High School[South]	30122.89	14855.15	2.03	0.05**	1.78

RSquare = 0.7532, RSquare Adjusted = 0.7158, F = 20.098, P(F > 20.098) <0.0001*
Note: *,**, and *** denote the statistical significance at 10%, 5%, and 1% levels

Our model estimates that the coefficient of slope for bedrooms is 14640.57. It means that, on average, we would expect the price to increase by $14,640.57 for each bedroom in the house, keeping the other variables fixed. The other coefficients can be interpreted in similar fashion.

The coefficients of slope for the dummy variables (High Schools) can be interpreted similarly and are more interesting in this example. The coefficient of slope for High School [Central] tells us that, on average, we expect that the price of a house in the Forsyth Central High School region is $15,791.96 less than a house in the West Forsyth high School region, keeping the other variables fixed. The coefficient of slope for High School [Denmark] tells us that, on average, we expect that the price of a house in the Denmark High School region is $5,544.18 higher than a house in the West Forsyth High School region, keeping the other variables fixed. The coefficient of slope for High School [Lambert] tells us that, on average, we expect that the price of a house in the Lambert High School region is $78,093.97 higher than a house in the West Forsyth high School region, keeping the other variables fixed. The coefficient of slope for High School [North] tells us that, on average, we expect that the price of house in the North Forsyth High School region is $56,409.39 less than a house in the West Forsyth high School region, keeping the other variables fixed. The coefficient of slope for High School [South] tells us that, on average, we expect that the price of a house in the South Forsyth High School region is $30,122.89 higher than a house in the West Forsyth High School region, keeping the other variables fixed. These interpretations are consistent with our experience that Lambert, Denmark, and the South Forsyth are the most popular high schools in the county.

CONCLUSIONS AND DISCUSSION

This paper has analyzed the drivers of housing prices for the city of Cumming, Georgia, one of the fastest-growing regions of the state, and in fact, of the United States, for the year 2020. More specifically, we studied the linear relationship of housing prices with number of bedrooms, number of bathrooms, square footage, lot size, number of garages, status of the house (active or not), and the high schools assigned to the houses. Data on 92 single family houses were collected and analyzed.

Our model suggests that an additional bathroom increases the price by $22,233.07 on average, while, an additional bedroom increases the price by $14640.57, on average, keeping the other variables fixed. This might sound contradictory to the common notion that an increase in the number of bedrooms increases the price of a house more than an increase in the number of bathrooms. However, this result is consistent with the result of Sirmans and Macpherson (2003), who found that adding a bathroom increased the sale price of a home by 8.7 percent, more than twice the rate for adding a bedroom. Of particular interest is the impact of the high school for which a house is zoned. The coefficients for Lambert, Denmark, and South Forsyth were positive compared to the reference category of West Forsyth. Further research reveals potential explanations. Denmark, opened in 2018, is the newest school in Forsyth County, possibly explaining why being zoned for

Denmark results in a positive model coefficient. Similarly, Lambert, opened during the 2009-2010 school year, is also a relatively new school in the county. South Forsyth houses the only International Baccalaureate program in Forsyth county, possibly explaining its positive contribution to price.

REFERENCES

2020 Best Counties for Families in America. (2020). Retrieved from https://www.niche.com/places-to-live/search/best-counties-for-families/

Anderson, D., Shoesmith, E., Sweeney, D., Anderson, D., & Williams, T. A. (2014). *Statistics for business and economics 3e*. Cengage.

Belke, A., & Keil, J. (2018). Fundamental determinants of real estate prices: A panel study of German regions. *International Advances in Economic Research, 24,* 25-45.

Bowerman, B. L., O'Connell, R. T., & Murphree, E. (2014). *Business statistics in practice*. McGraw-Hill/Irwin.

Bui, T. (2020). A study of factors influencing the price of apartments: Evidence from Vietnam. *Management Science Letters*, *10*(10), 2287-2292.

Calmasur, G. (2016). Determining factors affecting housing prices in Turkey with hedonic pricing model. International Conference on Business and Economics Studies, Washington D.C., USA, 255-269.

Cebula, R. J. (2009). The hedonic pricing model applied to the housing market of the city of Savannah and its Savannah historic landmark district. *The Review of Regional Studies*, *39*(1), 9-22.

Dubin, R. A. (1998). Predicting house prices using multiple listings data. *The Journal of Real Estate Finance and Economics*, *17*(1), 35-59.

Kain, J. F., & Quigley, J. M. (1970). Measuring the value of housing quality. *Journal of the American Statistical Association*, *65*(330), 532-548.

New Census Bureau Estimates Show Counties in South and West Lead Nation in Population Growth. (April 18, 2019). Retrieved from https://www.census.gov/newsroom/press-releases/2019/estimates-county-metro.html

Newsome, B. A., & Zeitz, J. (1992). Adjusting comparable sales using multiple regression analysis - The need for segmentation. *The Appraisal Journal*, 129-135.

Pardoe, I. (2008). Modeling home prices using realtor data. *Journal of Statistics Education*, *16*(2).

Randeniya, T. D., Ranasinghe, G., & Amarawickrama, S. (2017). A model to estimate the implicit values of housing attributes by applying the hedonic pricing method. *International Journal of Built Environment and Sustainability, IJBES, 4*(2), 113-120.

Selim, H. (2009). Determinants of house prices in Turkey: Hedonic regression versus artificial neural network. *Expert Systems with Applications*, *36,* 2843-2852

Sirmans, G. S., MacDonald, L., Macpherson, D. A., & Zietz, E. N. (2006). The value of housing characteristics: A meta-analysis. *The Journal of Real Estate Finance and Economics*, *33*(3), 215-240.

Sharpe, N. R., De Veaux, R. D., & Velleman, P. F. (2012). *Business statistics*. Pearson Education.

Wooldridge, M. (2011). *Modern econometrics*. McGraw Hill.

Yusof, A., & Ismail, S. (2012). Multiple regressions in analyzing house price variations. *Communications of the IBIMA*, 1-9.